

Process-Oriented Measurement Using Electronic Tangibles

Jochanan Veerbeek¹, Janneke Verhaegh², Julian G. Elliott³ & Wilma C. M. Resing¹

¹ Developmental and Educational Psychology, Leiden University, Leiden, The Netherlands

² Philips Research, Eindhoven, The Netherlands

³ School of Education, Durham University, Durham, UK

Correspondence: Wilma C. M. Resing, Faculty of Social Sciences, Department of Psychology, Section Developmental and Educational Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands. Tel: 31-71-527-3680. E-mail: resing@fsw.leidenuniv.nl

Received: November 22, 2016

Accepted: December 29, 2016

Online Published: January 10, 2017

doi:10.5539/jel.v6n2p155

URL: <http://dx.doi.org/10.5539/jel.v6n2p155>

Abstract

This study evaluated a new measure for analyzing the process of children's problem solving in a series completion task. This measure focused on a process that we entitled the *Grouping of Answer Pieces* (GAP) that was employed to provide information on problem representation and restructuring. The task was conducted using an electronic tangible interface, to allow for both natural manipulation of physical materials by the children, and computer monitoring of the process. The task was administered to 88 primary school children from grade 2 ($M=8.2$ years, $SD=0.50$). GAP was a moderate predictor of accuracy on the series completion task. Averaged over multiple items, GAP, verbalizations and time measures were related to accuracy. On an item level, however, GAP was the only process measure related to item solving success, and this relationship was mediated by item difficulty. Further research is needed to investigate the precise relationship between problem solving and GAP.

Keywords: cognitive strategy use, problem solving, inductive reasoning, process assessment, problem space restructuring, problem representation

1. Introduction

Throughout their school careers, children are subjected to a host of assessment procedures that seek to monitor their learning. In school, their cognitive and curricular progress is monitored by achievement tests; outside of the classroom, intelligence tests are sometimes used to assess the child's cognitive ability. However, these instruments have been subject to critique on the grounds that they are unable to provide information on how a child learns, offer few details about why the child failed to learn (Elliott, Grigorenko, & Resing, 2010; Elliott, 2000), and do not yield useful information about what forms of educational intervention might help the child (Elliott, 2000).

An alternative approach to cognitive assessment involves process-oriented measurement (Benson, Hulac, & Bernstein, 2013; Resing & Elliott, 2011). This form of measurement focuses on the *process* of problem solving, instead of, or in addition to, its *products*, and may help to explain why a particular child failed to solve a problem. Studying the operation of cognitive processes within a test situation could potentially yield information that aids the design of subsequent instruction and intervention (Elliott, 2000; Greiff et al., 2013; Van Gog, Kester, Nieuvelstein, Giesbers, & Paas, 2009).

In line with such reasoning, the general aim of the present study was to examine a new process measure that we called *Grouping of Answer Pieces* (GAP), which was designed to assess problem representation and restructuring. This measure was evaluated on its predictive properties within problem solving both in itself, and in combination with existing measures of the problem-solving process. Finally, this study aimed to evaluate the usefulness of a combination of an electronic tangible interface and dedicated analysis system in process-oriented assessment within a problem-solving framework.

1.1 The Process of Problem Solving

How people solve problems has been a major concern within the fields of cognitive psychology and artificial intelligence. While formal intelligence tests have been widely used, these have not proven very successful in helping us understand individual differences in particular problem-solving processes (Richard & Zamani, 2003).

The process of problem solving has often been described as cyclical, consisting of (1) problem recognition, (2) definition and representation of the problem, (3) development of a solution strategy, (4) organization of relevant knowledge, (5) allocation of mental and physical resources, (6) monitoring the progress towards the goal, and finally, (7) evaluation of the accuracy of the solution (Pretz, Naples, & Sternberg, 2003).

The second phase of the problem-solving cycle, the definition and representation phase, has been extensively described by Newell and Simon (1972), who introduced the concept of a problem space, indicating all possible solutions to the problem. According to these authors, problem space can be reduced by breaking down a problem into a set of smaller problems. Here, heuristics can serve as rules that determine how the problem can be divided into a series of smaller problems, leading to a restructuring of the problem space (Pretz et al., 2003). Heuristics are seen as fast rules and procedures for obtaining an answer or decision without the use of an algorithm, and, therefore, do not necessarily always lead to the correct or optimal result (Colman, 2006). Problem-solving strategies and problem representation are thought to influence each other, as both are related to problem-solving performance, as well as to transfer (Alibali, Phillips, & Fischer, 2009).

1.2 Measuring the Process of Problem Solving

The literature on the measurement of problem-solving processes has mainly focused on strategy use. A cognitive strategy has been defined by Kossowska and Ncka (1994) as “*a unique pattern of information-processing which takes place in a problem solving situation*” (p. 33). It is considered to be a vital component of the problem-solving process (Richard & Zamani, 2003; Siegler, 2007) having both an impact on, and being impacted by, learning (Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Siegler, 2004). Siegler (1996) described high quality learning as not being rigidly connected to one particular strategy, but rather, to one's ability to adapt strategy use flexibly to the task requirements. In his opinion, variability in strategy use (rather than stable use of one particular strategy) is often indicative of adaptive learning.

Several methods for measuring strategy use are available although each offers a compromise between accuracy of measurement, participant involvement and reactivity, and ease of use (Tenison, Fincham, & Anderson, 2014). Reactivity is understood here as a change in strategy use, as a result of its assessment. In such instances, the observed strategy use is different to how the participant might otherwise have employed strategies (Kirk & Ashcraft, 2001; Tenison et al., 2014).

Verbal (i.e., oral) reports are widely employed means of assessing strategy use and cognitive processes. Kirk and Ashcraft (2001) suggested that offering a verbal report may influence the natural mental processes of the participant (i.e., inducing reactivity). Such influence could, in theory, lead to improved or reduced performance. Verbal reporting might increase cognitive load demands and, as a result, reduce the mental resources available for the process that is being reported upon. On the other hand, participants might be motivated to work through the task with greater energy and accuracy, as the requirement to offer oral reports might expose their possible errors more publicly. Strategy assessment using verbal reports therefore requires a compromise between report accuracy and participant reactivity (Tenison et al., 2014). Although debate exists about the accuracy and reliability of verbal reports of cognitive processes (e.g., Feldon, 2010), the general consensus is that these provide valid data when obtained under the correct circumstances (Ericsson & Simon, 1980; Tenison et al., 2014). However, as noted above, we must recognize that the requirement to describe their use of cognitive processes may elicit reactivity from the participants, affecting their natural strategy use (Tenison et al., 2014).

Problem-solving speed has generally been presented in the literature as indicative of cognitive ability; the general assumption being that faster is better, although research findings have not unilaterally supported this view (Goldhammer et al., 2014; Scherer, Greiff, & Hautamäki, 2015). High performing participants have tended to be faster than less proficient participants at highly perceptual, automated, low complexity tasks, but they may take more time when tackling more challenging and complex reasoning tasks (Goldhammer et al., 2014). Others (e.g., Kossowska & Ncka, 1994) have examined the amount of time taken at different stages of task completion. By analyzing the proportion of time spent on the initial stages of the task, an estimate can be made of the portion of time a participant spent on the analysis of the task and their planning of the problem-solving process. Higher performing participants have been found to spend relatively more time than weaker performers on analysis and planning in the initial stages of the task (Kossowska & Ncka, 1994; Resing & Elliott, 2011; Resing et al., 2012).

Although developments in the field of technology have offered new possibilities for studying strategy use (Ericsson, 2003), advances in process-oriented measurement have yet to lead to widely used practical methods to incorporate process measures into the assessment of learning and cognitive abilities.

1.3 Inductive Reasoning

Inductive reasoning requires the detection of a rule governing a specific set of elements, and the formulation of a general rule from these elements (Klauer & Phye, 2008), in such a way that reasoning from a particular situation is applied to a general situation (Sternberg, 1985). Inductive reasoning is generally seen as important for learning and transfer (Klauer, Willmes, & Phye, 2002; Resing et al., 2012). Some authors have argued that inductive reasoning is an important component of cross-curricular thinking and learning skills (Greiff et al., 2013; Molnár, Greiff, & Csapó, 2013). A number of different types of tasks are based on the principles of inductive reasoning; these include analogies, series completion, and categorization (Sternberg, 1985). The focus of this paper is on series completion tasks, which require the solver to analyze a series of elements, and complete the series by supplying the missing element(s). Series completion problems exist in a number of shapes and forms, using letters, numbers, geometric figures, colours, etc. Some forms, such as letters and numbers, have a fixed relationship to each other as they have a natural sequence, while others, such as geometric figures and colours do not. Series completion has long been the subject of research, and the processes involved have been described extensively (see Holzman, Pellegrino, & Glaser, 1983; Simon & Kotovsky, 1963; Sternberg, 1985).

1.4 Electronic Tangibles

In studying the process of problem solving, computers are useful tools for registering test performance. Over the last years, computerized forms of intelligence tests have been introduced. An individual's performance on the Wechsler Intelligence Scale for Children®- Fifth Edition (WISC-V) can be recorded either on paper or using an iPad. Although offering a computerized version, such tools are not designed to measure the process of problem solving; the focus here is still on recording the outcome.

For those who seek to measure problem-solving processes, physical objects offer benefits that the traditional PC or tablet based interfaces would appear to lack. The benefits of physical materials for learning have been advocated by seminal writers such as Piaget, Bruner and Montessori whose theories inspired the development of sets of materials for classroom use. Dienes' multi-base arithmetic blocks (Dienes, 1964), for example, were intended to facilitate comprehension of elementary mathematics by the formation of "qualitative structures", such as the concept of number (e.g., Piaget, 1976). Digitized physical learning materials were introduced by Papert and his student Resnick, who developed so-called "digital manipulatives" (Resnick et al., 1998). Operating across a broader context than schooling alone, the concept of Tangible User Interfaces (TUIs) offers great promise for learning and assessment. TUIs consist of electronically enhanced tangible materials, which permit a seemingly more natural performance by the student and enable the collection and analysis of computer data by an assessor. In contrast to PC or touch-surface tablet applications, where 2 or 3 dimensional representations of objects are typically utilized, electronic tangibles make use of real objects (Verhaegh, Resing, Jacobs, & Fontijn, 2009).

TUIs integrate input and output in physical objects that represent digital information themselves (Ullmer & Ishii, 2000). Graphical User Interfaces (GUIs), such as a PC mouse and a screen, separate input and output modalities, whereas TUIs seamlessly integrate control and representation. Where younger children may experience difficulty performing some actions on touch-surface tablets, such as drag-and-drop procedures (Price, Jewitt, & Crescenzi, 2015), the physical materials that are used in TUIs permit more natural interaction with the interface (Verhaegh et al., 2009), and draw upon the use of a wider range of human skills and abilities such as perception, motor skills and emotion (Dourish, 2004). It has been found that early cognitive development depends mostly on sensory-motor responses (Goswami, 2008), and, thus, the use of tangible interfaces comes naturally to people.

Several possible benefits of TUIs for learning have been described. TUIs are assumed to support playful learning, which enhances children's engagement in scholastic learning tasks. Furthermore, it is likely that they offer a more accessible and direct interface than PC or Mac-based learning applications, and support multisensory learning as well as collaborative play (Manches, O'Malley, & Benford, 2009; Marshall, 2007).

1.5 Aims and Research Questions

The current research concerned an examination of a novel method of process-oriented measurement involving a new measure of strategy use, called *Grouping of Answer Pieces* (GAP). This measure was applied in a series completion construction task with a TUI, an electronic console. The task consisted of puppet figures, which were to be constructed using eight separate pieces. GAP was considered to be indicative of the use of adaptive heuristics, employed to reduce and (re)structure the problem space, and considered to represent the smaller problems that the task had been broken into (Pretz et al., 2003).

GAP in the series completion task was expected to be related to the accuracy of the participants' performance (Richard & Zamani, 2003). However, it was not expected to be related to other measures of strategy use, such as time measures or verbal reports, as these take place during different stages of the problem-solving cycle (Pretz et al., 2003). GAP was thought to add unique predictive and explanatory value to performance on the tangible series completion task, a factor that could hopefully be added to the existing array of measures, such as verbalizations of the problem-solving process, time measurement, and previous inductive reasoning ability.

Variability in strategy use between items was expected to be connected to performance, thus providing additional value in predicting test performance on the tangible series completion task. Participants who displayed greater variability in strategy use between items, were expected to perform better than those showing less variability (Siegler, 1996, 2007).

Finally, we anticipated that, for each item, performance on the tangible series completion task could largely be explained by a combination of measures: initial skill level, GAP, verbal reports, time strategies, and task features. Siegler (1987) has pointed out, however, that averaging data over multiple items can lead to a distorted image of an individual's strategy use, and can result in the loss of valuable information. Additional analysis of each item was expected to prevent the loss of information that could result from using data averaged over multiple items. The study sought to study the relationship between the process measures, task identity, previous ability, and performance on the tangible series completion task. As was found with the relationship between time measures and performance (Goldhammer et al., 2014; Scherer et al., 2015), we expected this relationship to be complex and interactive.

2. Method

2.1 Participants

The participants in this study were N=88 children, 46 boys and 42 girls (M=8.2 years; SD=0.50), from 4 grade 2 classes of 3 primary schools. The schools were selected on the basis of their willingness to cooperate and were all located in a predominantly middle class area in the Netherlands. Informed consent from the parents was obtained before testing. Three children failed to complete the study due to absence as a result of illness and were excluded from the analyses.

2.2 Design and Procedure

Participants were first presented with the Raven's Standard Progressive Matrices (Raven, Raven, & Court, 1998). Each participant received his/her own booklet and answer sheet, and was required to complete the matrices independently in the classroom. After the matrices had been completed, each child was taken out of class to work on the tangible series completion task individually on the electronic console. The console provided standardized instruction for all of the participants. An examiner was present at all times to collect and return the children to class, and to oversee the task process. However, they had no role in providing test instructions.

2.3 Materials

2.3.1 Raven's Standard Progressive Matrices

The Raven's Standard Progressive Matrices Test (Raven et al., 1998) was used to assess initial cognitive ability. This group test is considered to be a sound indicator of inductive reasoning ability.

2.3.2 Series Completion Task

This study used a schematic-picture inductive reasoning series completion task which was designed specifically for use with the TUI system, although initially designed as a dynamic test incorporating a graduated prompts form of training (e.g., Resing & Elliott, 2011; Resing, Touw, Veerbeek, & Elliott, 2016). The series completion task required the child to detect changes in objects and relationships in a series of puppet figures, and formulate a rule to complete the series. The task, based on the puppet series completion task designed by Resing and Elliott (2011) was intended to provide an indication of each child's inductive reasoning ability. Schematic-picture tasks such as the tangible series completion task used in this research are seen as more complex than series completion tasks that make use of letters or numbers. While letters and numbers have a fixed relationship to each other, pictures and colours do not. Thus, in order to solve the series, one must first search for repeating combinations of pictures prior to being able to understand the relationships between the elements of the task.

The test consisted of 12 items with increasing levels of difficulty. The test started with an example item. If the child was unable to provide the correct answer on the example item, the console would provide additional explanation to the child, to ensure understanding of what was expected of him/her. Each series item consisted of an initial array of six puppets and the child was asked to complete the sequence by making the seventh puppet.

The child had to analyze the changes across successive puppets and find the rule to enable them to complete the task. An example item can be found in Figure 1. Each puppet consisted of 8 separate pieces. The head was a single piece that determined the gender of the puppet (either boy or girl). The 7 pieces that made up the body, arms and legs of the puppet could vary in colour and pattern. There were 4 different colours available (green, blue, pink, and yellow), which could be plain (no pattern), dotted or striped. The design of the eight piece puppet allows for multiple transformations in a series, so the participant is called upon to use a large number of rules in order to complete the tasks.



Figure 1. An example of an item from the puppet series completion task. The child is presented with an array of puppets (drawn on paper) and is asked to construct the puppet that should appear next in the series

2.3.3 Electronic Tangible Interface

Our study employed an electronic console called “TagTiles” (Serious Toys, 2011), which enabled us to use a computerized environment for assessment, without the issues that manipulation of a touchscreen or mouse bring for young children (Price et al., 2015; Verhaegh et al., 2009). This incorporated a 12x12 electronic grid, which was equipped with sensors to detect the placement of puppet pieces on its surface, and LEDs which could be programmed to provide visual, brightly coloured feedback. Through its audio output, the console was able to provide appropriate task instructions. The series completion task was completed by placing the pieces on the console. Each contained a unique RFID tag, which enabled the sensors to detect position, timing, and identity of that particular piece on the console’s surface. All activity data were automatically saved in log files on SD memory cards. Log files contained rudimentary information about time, identity, and position of pieces placed on the console, and details about the accuracy the answers, per piece and for the item as a whole. The log files that were created by the console were manually cleared of unnecessary data, e.g., accidental movement of pieces, and relevant data were transferred into SPSS for analysis. Wherever possible, missing data, caused by any failure to detect pieces by the console, were retrieved from written records of the child’s performance made by the tester during testing.

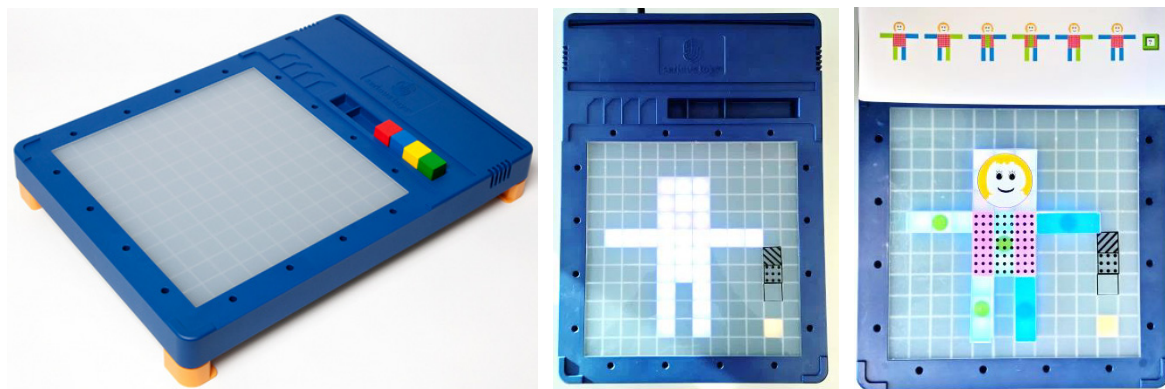


Figure 2. TagTiles console

2.4 Scoring and Analysis

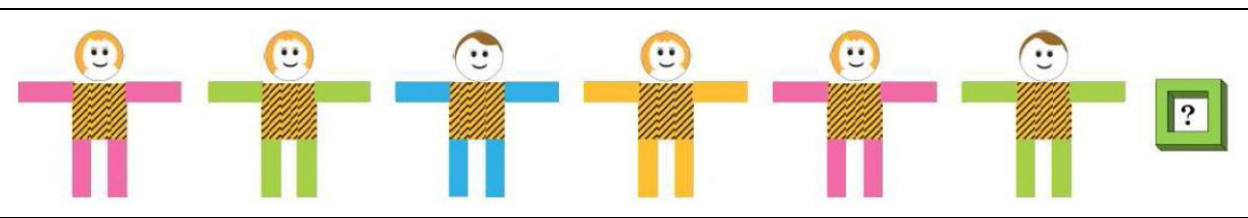
2.4.1 Accuracy

Accuracy was used as the primary outcome variable for the series completion task. The scoring was based on the number of correctly placed pieces. As each answer contained eight puppet pieces, the score for each item ranged from 0 to 8. Given that the full test consisted of 12 items, each child's accuracy score could range between 0 and 96. This approach was expected to enable far greater differentiation than if we had employed more straightforward right or wrong scoring for each item.

2.4.2 Grouping of Answer Pieces (GAP)

The concept of grouping in respect of the placement of the answer pieces on the task is similar to the principle of grouping into “chunks” that is widely utilized in memory contexts (Miller, 1994; Simon, 1974). This is based on the structuring of the problem space and the creation of sub-goals or groups by means of adaptive heuristics. As the puppets used in the tangible series completion task are composed of multiple pieces, and the series completion tasks are composed of a number of different relations and transformations, the child's response sequence is thought to be influenced by the unfolding process of solving the tasks. The subdivisions in the sequence of piece placement were theorized as indicative of the concepts used to define and represent the problem (Pretz et al., 2003).

Task characteristics were used to create sub-groups of puppet pieces that go through the same transformation, or which are grouped on the basis of colour, pattern, or anatomy. Puppet pieces were considered to be grouped if they were placed immediately after each other. First, puppet pieces were numbered, so a sequence could be identified indicating which piece was placed at a particular point in time. The identification numbers of the pieces ranged from one to eight, as follows: (1) head, (2) left arm, (3) right arm, (4) left body, (5) middle body, (6) right body, (7) left leg, and (8) right leg. An example item (Figure 3) illustrates the basic principle of the GAP measure. The Figure includes a sequence of puppets (the task presented to the child) illustrated with some possible sequences of responses. The displayed task consists of three discernible series of transformations. First, the heads go through a series of changes. The second series is that of the arms and legs, which change colour. The third series is that of the body, which stays the same throughout the series. The first sequence displayed in Figure 3 contains no adaptive grouping of answer pieces, as each successive piece placed is one that goes through a different series of transformations. No adaptive groups of puppet pieces were constructed; pieces that transform according to the same rule were not grouped together. Example 2 contains one of the two groups, the “body” group (the 4th, 5th and 6th position, pieces 4, 5 and 6). All the pieces in this group are placed in immediate succession to one other. The “arms+legs” group (pieces 2, 3, 7 and 8) were not placed as a group, as they were interrupted by the body group. Finally, Example 3 contains both the “body” and the “arms+legs” groups. Here, all pieces of both groups were placed as a group following each other in the sequence.



Groups for this item		1st	2 nd	3rd	4th	5th	6th	7th	8 th
- Body (4 5 6)	Example 1	1	2	4	3	5	7	6	8
- Arms + Legs (2 3 7 8)	Example 2	1	2	8	6	5	4	7	3
	Example 3	1	5	4	6	2	3	7	8

Figure 3. Illustrative responses showing grouping strategy (GAP). Three sequences are provided, with the 1st position being the first piece placed in the sequence, up to the 8th and last piece. The groups for this item are listed in the far left column. Heads are treated as a separate piece and are not included in any of the groups

The adaptive groups of puppet pieces that could be utilized differed between the various items, with the number and type of groups that could be discerned per item ranging between two and five. Some items contained groups that overlapped or contained other smaller groups. For each test item, formulae were written in Microsoft Excel

to identify the placement of adaptive groups for a particular item. The GAP score was based on the number of groups laid down by the participant, divided by the number of possible groups for that item.

2.4.3 Verbalized Strategies

After each item was completed, the console's electronic voice asked "*Why do you think this is the correct puppet?*". Children's answers were recorded in writing and by audio recordings, and the verbalizations of their solution strategies were recorded and scored. The scoring system was based on verbalizations that had been found in previous studies, the literature on strategy use in inductive reasoning tasks, and categories used in prior research on reasoning tasks. This resulted in three levels of verbalized strategies, as used by Resing and colleagues (2016), and depicted in Figure 4. In the first group (I), children were able to provide full explanations of all transformations involved in the series, either explicitly or implicitly (e.g., by pointing). The second group (II) contained children who were able to verbalize some transformations in the puppet series, but not all those that were needed to solve the task. The third group (III) consisted of verbalizations that did not provide information relevant to the solution of the task. This scoring was used to analyze each item. The children were then allocated to 1 of 5 classes of their verbalizations. If children used a single type of verbalization for more than 33% of the items, they were allocated to the corresponding strategy class. If they used two types of verbalizations, both in more than 33% of the items, they were assigned to a mixed strategy class (Figure 4).

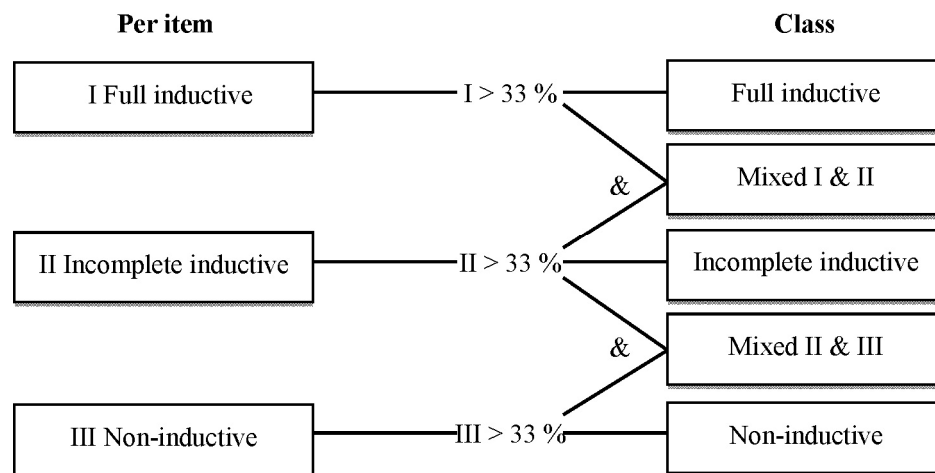


Figure 4. Verbalized strategies. Verbalization is scored for each item. The children were assigned to one of five verbalized strategy classes on the basis of the percentage of items where a particular type of verbalization was provided

2.4.4 Time Strategies

Time measures obtained from the log files were the total time for completion of the task (I_{TOTAL} , Figure 5) and the time intervals between the placement of all pieces.

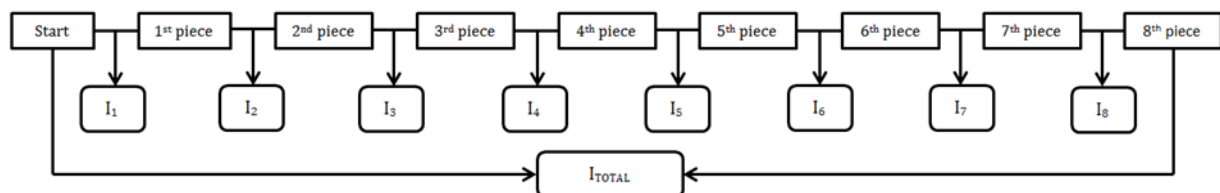


Figure 5. Interval calculation

An adapted version of Kossowska and Ncka's (1994) formula was used to calculate the proportion of time used on the initial stages of the problem-solving process. These stages are considered to reflect the time taken for analysis and planning of the problem-solving process (Kossowska & Ncka, 1994; Resing & Elliott, 2011; Resing

et al., 2012). The formula calculated the proportion of time used for the placement of the first two pieces, as previous research has shown that many children place the puppet head first and then progress to completing the rest of the puppet. This resulted in the following formula:

$$\text{ThinkingTime} = \frac{I_1 + I_2}{I_{\text{TOTAL}}} \quad (1)$$

Higher values for thinking time represented more thinking and planning in advance; lower values indicated a more impulsive style of addressing the task (Resing & Elliott, 2011).

2.4.5 Decision Tree Analysis

Classical linear analyses are widely used to investigate contributory predictive factors. However, their usefulness is limited when they are used with data that contain complex interactions and which are non-linear (Ritschard, 2014). Decision Tree Analysis (DTA) contains a number of exploratory techniques aimed at detecting interactions and non-linear relationships within a dataset. The basis of DTA is recursive repartitioning, which involves splitting the data in order to achieve the optimal difference between groups on the outcome variable. This procedure is repeated for each of the splits until an appropriate place to stop is reached. DTAs compare all predictors and search through all possible cut-off points with respect to their effect on the outcome variable. The splitting variable is chosen as a predictor that maximizes the relationship with the outcome variable, at a specific cut-off point (McArdle, 2014). The DTA framework offers multiple techniques and forms of statistical analysis.

We employed the CHAID technique in the present study (McArdle, 2014; Ritschard, 2014). CHAID uses Chi-square analysis as its splitting criterion. The advantage of this approach is that it permits splitting into more than two groups at once for a single predictor. These groups of cases resulting from a split are called “Nodes” in DTA. At each splitting point, CHAID determines the optimal number of splits, and the cut-off points for these splits, for each of the predictors. The p-value of the Chi-square (with a Bonferroni correction) is used as the criterion to determine the splits. The p-values are sensitive to the number of cases involved, and help avoid any splitting into groups that are too small. The minimum number of cases involved in each split can also be predetermined (Ritschard, 2014), as was done in this study.

3. Results

3.1 Validating GAP on a Test Level

Firstly, we examined whether the GAP score was correlated with performance on the task. As expected, we found a positive (albeit moderate) correlation between GAP and Accuracy ($r=.32, p=.002$). Additionally, and also as expected, no significant correlations were found for Verbalized strategies ($r=.02, p=.88$), TotalTime ($r=-.05, p=.65$), or ThinkingTime ($r=.17, p=.12$) with GAP. These findings indicate that GAP can be seen as a unique measure of strategy use, unrelated to any previous measures of strategy use.

Multiple regression analysis was used to investigate the prediction of Accuracy in completing the task. Multiple models were tested, and the results are depicted in Table 1. In the first model, Accuracy was used as the dependent variable and GAP was used as the independent variable. As the sole predictor, GAP explained 9.0% of the variance in Accuracy. In Model 2, Verbalization, TotalTime and ThinkingTime were added to the model as independent variables. This model explained 27.0 % of the variance in Accuracy. GAP, Verbalization and ThinkingTime were found to be significant predictors of Accuracy. In contrast, TotalTime was not a significant predictor. The final model (Model 3) contained Raven scores as an independent variable, along with the independent variables used in the previous model. This model explained 30.7% of the variance in Accuracy, with GAP, Verbalization and Raven scores as significant predictors for Accuracy. Neither TotalTime nor ThinkingTime were significant predictors of Accuracy. These findings were in line with our expectations that GAP would add unique predictive value to the available measures.

Table 1. Regression analysis for accuracy on the puppet task

Variable	Model 1			Model 2			Model 3		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	<i>B</i>
Constant	54.87	7.38		24.32	9.49		19.95	9.43	
GAP	33.72	10.91	.32**	29.83	10.02	.28**	26.40	9.88	.25**
Verbalization				3.11	1.07	.27**	3.21	1.04	.28**
TotalTime				0.00	0.00	.19	0.00	0.00	.16
ThinkingTime				40.84	17.00	.24*	29.20	17.31	.17
Raven scores							0.36	0.15	.23*
R^2		.10			.30			.35	
<i>F</i> for change in R^2		9.55**			7.96**			5.39*	

* $p < .05$. ** $p < .01$.

3.2 Variability in Strategy Use

The second hypothesis reflected our expectation that variability in strategy use would be indicative of superior performance on the puppet task, and also, of overall learning. To investigate this, we calculated the variance within each participant's strategy use across all of the items. A multiple regression analysis was used with Accuracy on the task as the dependent variable, and variance in GAP, Verbalization, TotalTime and ThinkingTime as the independent variables. The results are presented in Table 2. This analysis yielded a model that explained 15.1% of the variance in Accuracy. A closer look at the model identified variance in GAP and variance in Verbalization as significant predictors of Accuracy, although it should be noted that GAP was negatively related to Accuracy. Although this was partly in line with our expectations, we had not anticipated the finding that more variability in GAP would lead to less accuracy on the task itself.

Table 2. Regression analysis with variability in strategy use

Variable	Accuracy		
	<i>B</i>	<i>SE B</i>	β
Constant	76.79	4.86	
GAP	-72.70	34.69	-.21*
Verbalization	31.17	10.35	.31**
TotalTime	-2.06E-10	0.00	-.01
ThinkingTime	163.25	150.18	.12
R^2		.19	
<i>F</i> for change in R^2		4.23**	

* $p < .05$. ** $p < .01$.

3.3 Validating GAP on an Item Level

The relationship between the number of correctly placed pieces on the puppet task items, the task characteristics, and the process measures was expected to be complex and interactive. To investigate this, a file was created in which each individual item for a particular participant was handled as a separate case ($N=1056$). For predicting Accuracy on each item (the number of body parts correctly placed out of eight), a Classification Tree was generated, using the CHAID method. The results of this analysis are displayed in Figure 6. Accuracy was used as the dependent variable, and the Item number (the number of the item on the tangible series completion task),

scores on the Raven's progressive matrices, TotalTime, ThinkingTime, GAP and Verbalization were used as the independent variables. The minimum number of cases per node was set to N=50.

For Items 4, 6, 9, and 10, the item number was the only factor to explain Accuracy. For Items 1 and 3, Accuracy was further explained by the Raven scores. For Item 2, 5, 7, 8, 11 and 12, GAP was added as an indicator of task success. Higher GAP values seemed to be predictive of higher Accuracy, although Node 10 is an exception to this. Node 10 was split on the basis of Raven's scores, with higher scores predictive of higher Accuracy. TotalTime, ThinkingTime, and Verbalization failed to offer a significant improvement of the model. At the item level, the best predictors of Accuracy were found to be task characteristics, previous ability (Raven's scores), and GAP.

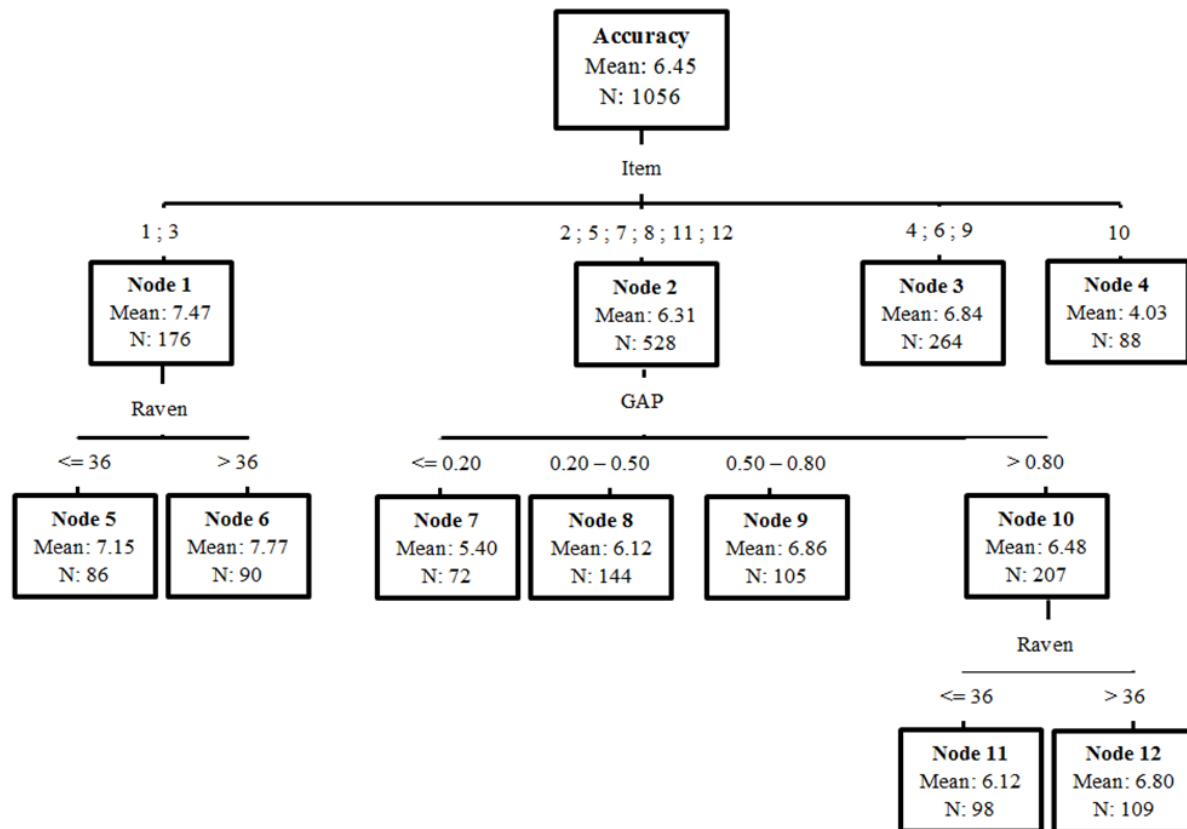


Figure 6. CHAID tree for the prediction of accuracy per item

4. Discussion

The aim of this paper was to gain greater understanding of how we can identify and assess the operation of problem-solving processes in young children. To assist in achieving this aim, we utilized a sophisticated assessment tool incorporating electronic tangible technology.

The use of the electronic tangible TagTiles console made it possible to observe and analyze children's problem-solving strategies in significant detail. GAP appeared to be moderately related to accuracy on the tangible series completion task, as were the previously available measures of strategy use, with the exception of TotalTime. This latter finding may be a consequence of the level of difficulty of the task, as time on task has been shown to be moderated by task difficulty (Goldhammer et al., 2014). The lack of a relation with the other measures suggests that GAP is a measure that can provide unique information about the process of an individual's problem solving.

Our findings regarding variability in strategy use showed a more complex picture. In line with Siegler's (1996) theory, intra-individual variability in both ThinkingTime and verbalizations was positively related to performance on the task. Variability in GAP was negatively related to accuracy; presumably more stable use of

grouping is indicative of greater accuracy in solving the task. If so, GAP appears to function differently to the other strategies. If GAP is more related to general task structuring, as we suspect, it would be a relatively constant style of approaching the process of problem solving, rather than a choice of strategy application. This would make the use of GAP less dependent on particular task content. As GAP is assumed to be related to problem representation (Pretz et al., 2003; Richard & Zamani, 2003), it may be a metacomponent of problem solving. Metacomponents, such as problem recognition, definition, and representation, were described by Sternberg (1985) as executive processes that guide the problem-solving processes, and were expected to be general across cognitive problem-solving activities (Pretz et al., 2003; Sternberg, 1985). As such, GAP may provide more general information on a child's problem-solving processes than previously available process measures, such as verbalizations and time measures, which were found to be more variable.

We were somewhat surprised to discover that the Items themselves were identified as the primary factor for task success as, at least superficially, these seem fairly similar in the characteristics and processes used. However, the results showed that task characteristics were important here. This finding builds upon the work of Goldhammer and colleagues (2014) who found that the relationship between time on task and performance was moderated by task difficulty. In other words, task difficulty is a key factor in determining the need for adequate strategy use. Analysis of the items in the present study indicated this to be true for our task. On the items where there was high average task success (in other words, where the item was found to be relatively easy), the item identity itself provided the best explanation for task performance. As Klauer and Phye (2008) have proposed, experts on a task may use less sophisticated strategies, requiring less time and effort, on easier items. Thus, the use of less sophisticated strategies on easier items may prove more efficient (Siegler, 1996). It was also seen that no single process-oriented measure provided an explanation for all items. This is in line with Siegler's (1996) notion that averaging data over multiple items might lead to the oversight of important information and may lead to an oversimplification of models of strategy use. GAP provided the single best explanation of all process-oriented measures for performance, as it was the only process-oriented measure included in the model. However, it was not a predictor for accuracy in all of the tangible series completion items. The relationship between strategy use and performance appeared to be moderated by task characteristics (Dodonova & Dodonov, 2013; Goldhammer et al., 2014; Tenison et al., 2014), and might be distorted by the use of linear analyses.

Although our two time measures offered additional explanatory value, no analysis or model included them both. Neither did they complement each other in the prediction of task performance, although ThinkingTime proved to have greater explanatory value. The finding that thinking and planning time provided more information than did total time is in line with the results of a study by Resing and Elliott (2011), who found that total completion time failed to discriminate between their trained and untrained participants.

4.1 Process-Oriented Measurement

Although process-oriented measurement yielded some additional explanatory value, a number of complexities emerged. In line with previous research, we found that process-oriented measures are dependent on task characteristics and do not show a unilateral relationship with performance (Goldhammer et al., 2014; Scherer et al., 2015). While process measures may provide additional information for the more difficult items, their relationship with easier items is unclear.

Process-oriented measurement is often labelled as strategy use in the literature. Such inconsistency in the use of terms may lead to confusion. Time measures, verbalizations and actions (GAP or otherwise) have all been labelled as strategy use in previous studies, but they all take place in, and over, different stages of the problem-solving cycle (Figure 7). Although the problem-solving cycle is not necessarily followed in a straightforward, linear fashion (Pretz et al., 2003), the different stages in the problem-solving cycle, where strategy measurement can take place, may explain why the different measures of strategy use were found to be unrelated.

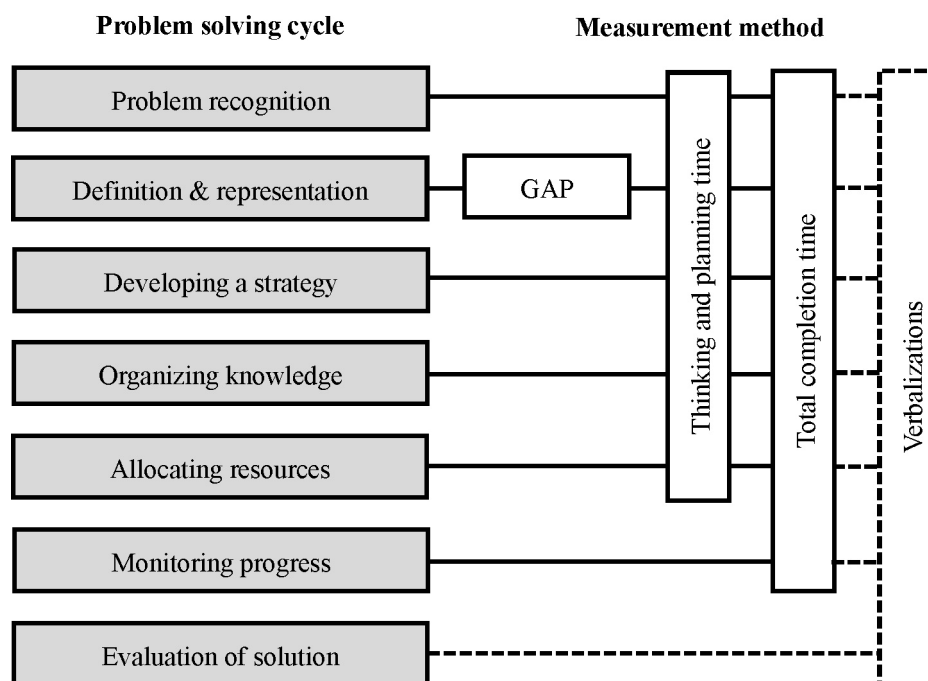


Figure 7. Different process-oriented measures with respect to the phases of the problem-solving cycle (Pretz et al., 2003)

GAP has a particular place in the problem-solving cycle, as it may be influenced by the phases between representation and taking physical action. Thinking time represents the time taken on the initial phases of the task, up to the point that physical action begins. Total completion time represents the time taken throughout the whole process, with the exception of the evaluation phase. Verbalizations about strategy use may concern any of the problem-solving phases, but would rarely cover all of these. The particular focus of such accounts will most likely be influenced by the type of verbalization measurement (Tenison et al., 2014), the accompanying instructions (Ericsson & Simon, 1980) and the participant's willingness and ability to reflect, report and elaborate upon his or her cognitive processes. Taking this into account, a more precise differentiation in the use of terminology could be desirable. We would suggest reserving the term *strategy* for any domain specific procedure, such as the math-specific strategies described by Siegler (1996). Differences in procedures aimed at general structuring of the problem-solving process, such as the time taken for planning and analysis (Kossowska & Ncka, 1994), or the grouping performed during problem solving, may be more accurately termed as *process structures*.

Although these findings offer some promising results with regard to process-oriented measurement using electronic tangibles, and the GAP measure in particular, caution is advised in interpreting these findings. As this research solely employed a tangible series completion task, our results cannot be generalized to other domains of problem solving. Even within the field of series completion, the particular task used in the present study cannot be readily generalized to other series completion tasks as these may not contain multiple transformations, or may be more domain-bound by the use of letters or numbers (Resing & Elliott, 2011).

It is also important to note that our sample size was rather small and spanned a narrow age range, thus further limiting generalizability. Clearly, more research is needed, utilizing larger and more diverse samples, in order to enable general statements to be made about the value of process-oriented measurement in education and clinical settings.

Although we found a relationship between performance and process measures in our particular test domain, its nature remains unclear. Future research should determine whether process training can be successfully employed to improve children's intellectual performance.

4.2 Conclusions and Recommendations for Future Research

In summary, our GAP procedure appears to offer additional and unique explanatory value to the field of process-oriented measurement. GAP can be measured and interpreted by technological systems such as that employed in the present study, and thus used in computerized testing environments where no examiner is present. Such a facility also makes the measure particularly suitable for analyses of classroom based problem solving. Although theoretically, it would be possible to measure and analyze such processes without the use of technology, this would not be recommended as the adaptive groups differ between items. Manual analysis would be prone to mistakes and oversights, and would be very time consuming. Tangible user interfaces, in contrast, make it possible for educators to provide individualized forms of adaptive instruction, based on the real-time activities/responses of the child.

Further research is needed to determine how to interpret the values derived from the obtained measures. In this respect, classification and regression trees may provide a more informative view than traditional analyses, as they are able to handle more complex and interactive data. This will enable researchers to take into account the interaction between item characteristics and the various process measures.

As GAP is based on the subdivision of tasks into sub goals, it is not possible to use this measure with all tasks. Future research should be aimed at identifying a range of diverse tasks that can enable measurement of this kind. The GAP measure does provide an opportunity for analyzing the problem-solving process of participants who for some reason are not able to provide (reliable) verbalizations of their strategies, such as those with specific language difficulties, certain children from ethnic minorities, etc. As the measurement is unobtrusive, there is no risk of participant reactivity. Additionally, tangible user interfaces offer the potential of providing individuals or groups of children with adaptive scaffolds, based on their differing responses to challenging classroom material. In that way, individualized training and assessment of the process of problem solving in education may come within reach.

Acknowledgments

The authors wish to thank Bart Vogelaar for his constructive comments on this paper and his support in preparing the final manuscript.

References

- Alibali, M. W., Phillips, K. M. O., & Fischer, A. D. (2009). Learning new problem-solving strategies leads to changes in problem representation. *Cognitive Development*, 24(2), 89-101. <https://dx.doi.org/10.1016/j.cogdev.2008.12.005>
- Benson, N., Hulac, D. M., & Bernstein, J. D. (2013). An independent confirmatory factor analysis of the Wechsler intelligence scale for children-fourth edition (WISC-IV) integrated: What do the process approach subtests measure? *Psychological Assessment*, 25(3), 692-705. <https://dx.doi.org/10.1037/a0032298>
- Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between South African and British students on the Wechsler Adult Intelligence Scales Third Edition (WAIS-III). *Frontiers in Psychology*, 6, 1-11. <https://dx.doi.org/10.3389/fpsyg.2015.00297>
- Colman, A. M. (2006). *Dictionary of Psychology*. New York: Oxford University Press.
- Dienes, Z. (1964). *Building up mathematics* (2nd ed.). London: Hutchinson Educational.
- Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence*, 41(1), 1-10. <https://dx.doi.org/10.1016/j.intell.2012.10.003>
- Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1), 19-30. <https://dx.doi.org/10.1007/s00779-003-0253-8>
- Elliott, J. G. (2000). The Psychological Assessment of Children with Learning Difficulties. *British Journal of Special Education*, 27(2), 59-66. <https://dx.doi.org/10.1111/1467-8527.00161>
- Elliott, J. G., Grigorenko, E. L., & Resing, W. C. M. (2010). Dynamic assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (Vol. 3, pp. 220-225). Oxford: Elsevier. <https://dx.doi.org/10.1016/B978-0-08-044894-7.00311-0>
- Ericsson, K. A. (2003). The acquisition of expert performance as problem solving. In J. E. Davidson, & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 31-83). New York: Cambridge University Press. <https://dx.doi.org/10.1017/CBO9780511615771.003>

- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215-251. <https://dx.doi.org/10.1037/0033-295X.87.3.215>
- Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence*, 35(4), 319-334. <https://dx.doi.org/10.1016/j.intell.2006.08.009>
- Feldon, D. F. (2010). Do psychology researchers tell it like it is? A microgenetic analysis of research strategies and self-report accuracy along a continuum of expertise. *Instructional Science*, 38(4), 395-415. <https://dx.doi.org/10.1007/s11251-008-9085-2>
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3), 608-626. <https://dx.doi.org/10.1037/a0034716>
- Goswami, U. (2008). *Cognitive development: The learning brain*. Hove, UK: Psychology Press.
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364-379. <https://dx.doi.org/10.1037/a0031856>
- Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, 75(4), 603-618. <https://dx.doi.org/10.1037/0022-0663.75.4.603>
- Kirk, E. P., & Ashcraft, M. H. (2001). Telling stories: The perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, 27(1), 157-175. <https://dx.doi.org/10.1037/0278-7393.27.1.157>
- Klauer, K. J., & Phye, G. D. (2008). Inductive Reasoning: A Training Approach. *Review of Educational Research*, 78(1), 85-123. <https://dx.doi.org/10.3102/0034654307313402>
- Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing Inductive Reasoning: Does It Transfer to Fluid Intelligence? *Contemporary Educational Psychology*, 27(1), 1-25. <https://dx.doi.org/10.1006/ceps.2001.1079>
- Kossowska, M., & Neka, E. (1994). Do it your own way: Cognitive strategies, intelligence, and personality. *Personality and Individual Differences*, 16(1), 33-46. [https://dx.doi.org/10.1016/0191-8869\(94\)90108-2](https://dx.doi.org/10.1016/0191-8869(94)90108-2)
- Manches, A., O'Malley, C., & Benford, S. (2009). The role of physical representations in solving number problems: A comparison of young children's use of physical and virtual materials. *Computers & Education*, 54(3), 622-640. <https://dx.doi.org/10.1016/j.compedu.2009.09.023>
- Marshall, P. (2007). Do tangible interfaces enhance learning? In *Proceedings of the 1st International Conference on Tangible and Embedded interaction* (pp. 163-170). New York: ACM Press. <https://dx.doi.org/10.1145/1226969.1227004>
- McArdle, J. J. (2014). Exploratory data mining using decision trees in the behavioral sciences. In J. J. McArdle, & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 3-47). New York: Routledge.
- Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 101(2), 343-352. <https://dx.doi.org/10.1037/0033-295X.101.2.343>
- Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity*, 9, 35-45. <https://dx.doi.org/10.1016/j.tsc.2013.03.002>
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Piaget, J. (1976). *To understand is to invent. The future of education*. New York: Penguin Books.
- Pretz, J. E., Naples, A. J., & Sternberg, R. J. (2003). Recognizing, defining, and representing problems. In J. E. Davidson, & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 3-30). New York: Cambridge University Press. <https://dx.doi.org/10.1017/CBO9780511615771.002>
- Price, S., Jewitt, C., & Crescenzi, L. (2015). The role of iPads in pre-school children's mark making development. *Computers & Education*, 87, 131-141. <https://dx.doi.org/10.1016/j.compedu.2015.04.003>

- Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Standard progressive matrices*. Oxford: Oxford Psychologists Press.
- Resing, W. C. M., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: Measuring children's change in strategy use with a series completion task. *The British Journal of Educational Psychology*, 81, 579-605. <https://dx.doi.org/10.1348/2044-8279.002006>
- Resing, W. C. M., Touw, K. W. J., Veerbeek, J., & Elliott, J. G. (2016). Progress in the inductive strategy use of children from different ethnic backgrounds: A study employing dynamic testing. *Educational Psychology*, 36. <https://dx.doi.org/10.1080/01443410.2016.1164300>
- Resing, W. C. M., Xenidou-Dervou, I., Steijn, W. M. P., & Elliott, J. G. (2012). A "picture" of children's potential for learning: Looking into strategy changes and working memory by dynamic testing. *Learning and Individual Differences*, 22(1), 144-150. <https://dx.doi.org/10.1016/j.lindif.2011.11.002>
- Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K., & Silverman, B. (1998). Digital manipulatives: New toys to think with. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 281-287) (CHI 98). New York: ACM Press/Addison-Wesley Publishing Co. <https://dx.doi.org/10.1145/274644.274684>
- Richard, J. F., & Zamani, M. (2003). A problem-solving model as a tool for analyzing adaptive behavior. In R. J. Sternberg, J. Lautrey, & T. I. Lubart (Eds.), *Models of Intelligence* (pp. 213-226). Washington: American Psychological Association.
- Ritschard, G. (2014). CHAID and earlier supervised tree methods. In J. J. McArdle, & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 48-74). New York: Routledge.
- Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the Relation between Time on Task and Ability in Complex Problem Solving. *Intelligence*, 48, 37-50. <https://dx.doi.org/10.1016/j.intell.2014.10.003>
- Serious Toys. (2011). *Serious Toys*. Retrieved from <http://www.serious toys.com>
- Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, 116(3), 250-264. <https://dx.doi.org/10.1037/0096-3445.116.3.250>
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.
- Siegler, R. S. (2004). Learning About Learning. *Merrill-Palmer Quarterly*, 50(3), 353-368. <https://dx.doi.org/10.1353/mpq.2004.0025>
- Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, 10(1), 104-109. <https://dx.doi.org/10.1111/j.1467-7687.2007.00571.x>
- Simon, H. A. (1974). How big is a chunk? *Science*, 183, 482-488. <https://dx.doi.org/10.1126/science.183.4124.482>
- Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, 70(6), 534-546. <https://dx.doi.org/10.1037/h0043901>
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.
- Tenison, C., Fincham, J. M., & Anderson, J. R. (2014). Detecting math problem solving strategies: An investigation into the use of retrospective self-reports, latency and fMRI data. *Neuropsychologia*, 54, 41-52. <https://dx.doi.org/10.1016/j.neuropsychologia.2013.12.011>
- Ullmer, B., & Ishii, H. (2000). Emerging frameworks for tangible user interfaces. *IBM Systems Journal*, 39, 915-931. <https://dx.doi.org/10.1147/sj.393.0915>
- Van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior*, 25(2), 325-331. <https://dx.doi.org/10.1016/j.chb.2008.12.021>
- Verhaegh, J., Resing, W. C. M., Jacobs, A. P. A., & Fontijn, W. F. J. (2009). Playing with blocks or with the computer? Solving complex visual-spatial reasoning tasks: Comparing children's performance on tangible and virtual puzzles. *Educational & Child Psychology*, 19, 18-39.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).